RAJAPINNOILLA-TEEMANUMERO

Datapaketit ja rajapinnat tutkijoiden työkaluna Kansalliskirjastossa: mahdollisuudet ja rajoitteet

Erno Liukkonen
Kansalliskirjasto, Helsingin yliopisto
erno.liukkonen@helsinki.fi
https://orcid.org/0000-0001-8611-951X

Liisa Näpärä
Kansalliskirjasto, Helsingin yliopisto
liisa.napara@helsinki.fi
https://orcid.org/0000-0002-6473-5896

Tuula Pääkkönen
Kansalliskirjasto, Helsingin yliopisto
tuula.paakkonen@helsinki.fi
https://orcid.org/0000-0003-3958-9732

Juha Rautiainen
Kansalliskirjasto, Helsingin yliopisto
juha.rautiainen@helsinki.fi
https://orcid.org/0000-0002-5223-9940

Pysyvä osoite: https://doi.org/10.23978/inf.107887

In this article, we study the possibilities of the data packages and interfaces of the National Library of Finland in the context of researcher use and comparing with other libraries and their solutions. The researchers' needs towards digital materials are various depending on the research field, usage purpose and the technical background each individual has. The researchers approach the digital materials from the viewpoint of their research needs, e.g. via source literature and materials. The National Library offers material and data from collections of the library, which have been digitized and converted to data packages based on the needs of the library and the researchers. The data packages and interfaces can be used e.g. in digital humanities, where one of the goals is to develop new methods and techniques. Presently the data packages and interfaces are, however, underused and seem hard to approach. Both the skills of the researchers, attitudes and understanding about interfaces and problems with interfaces limit their use. For example, the level of documentation, examples provided and visibility to the documentation on websites are not responding to the expectations of the researchers. In different libraries, the aim is to create tools, which aim to instruct and ease the use of the interfaces.

Asiasanat: digitaaliset ihmistieteet, sähköinen aineisto, rajapinnat (tietokoneohjelmat), dokumentointi, tutkijat, benchmarking, digitaalisten aineistojen tutkimuskäyttö, digitaalisten aineistojen kehitys, tutkijapalveluiden kehitys

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Johdanto

Kulttuuriperintöorganisaatiot, kuten kansalliskirjastot, ovat perinteisesti tärkeitä humanistiselle tutkimukselle. Niiden tarjoamat aineistot soveltuvat hyvin digitaalisten ihmistieteiden tutkimukseen, jossa yhdistyvät monitieteisesti ihmistieteiden kysymykset ja tietokoneavusteiset menetelmät (Schlicht, 2021, 26; Schwandt, 2021, 16). Digitaaliset menetelmät avaavat Kansalliskirjaston kokoelmien tutkimukselle uusia mahdollisuuksia, jos ne ovat tarjolla avoimena datana. (Candela et al., 2020, 2.) Kansalliskirjasto ja kansainvälisesti monet muut organisaatiot ovat luoneet digitaalisten aineistojensa yhteyteen datapaketteja ja rajapintoja, joiden avulla erilaiset tutkijat voivat hyödyntää niitä tietokoneavusteisin menetelmin.

Tutkijat lähestyvät kirjaston aineistoja ja dataa omien tutkimustarpeidensa ohjaamina, esimerkiksi lähdekirjallisuutta ja lähdeaineistoja tarvitessaan. Kansalliskirjasto puolestaan tarjoaa tutkijoille aineistoa ja dataa kirjaston kokoelmista, joita on vaihtelevasti digitoitu ja muunnettu datapaketeiksi

tutkijoiden käyttötarpeiden ja -pyyntöjen perusteella sekä muita periaatteita noudattaen. Lisäksi Kansalliskirjasto tarjoaa tutkimukseen sopivaa kuvailudataa, esimerkiksi kansallisbibliografia Fennican ja sanastopalvelu Finton kautta. Kansallisbibliografia on suomalaisen julkaisutuotannon tietokanta, joka sisältää tietoja mm. kirjoista, lehdistä ja sarjajulkaisuista. Finto on palvelu eri alojen yhteentoimiville sanastoille, ontologioille ja luokituksille. Käytössä on myös teknisiä rajapintoja, jotka avaavat mahdollisuuksia digitaalisten aineistojen tutkimukselle ohjelmallisesti, erityisesti digitaalisten ihmistieteiden kontekstissa.

Tässä artikkelissa keskityimme digitoiduista aineistoista tehtyihin datapaketteihin ja tutkijoille tarjottaviin rajapintoihin. Artikkelissa analysoimme datapakettien ja rajapintojen soveltuvuutta erilaisiin tutkimuksellisiin tarpeisiin hyödyntäen kansainvälistä vertailua ja oletettua tavoitetasoa, joka tukeutuu tulkintaan FAIR-periaatteista, eli löydettävyydestä (Findable), saavutettavuudesta (Accessible), yhteentoimivuudesta (Interoperable) ja uudelleenkäytettävyydestä. Artikkelin aineisto perustuu Digitaalinen avoin muisti (DAM)-hankkeessa tuotettuun aineistoon, joka on koottu useista tutkijoille suunnatuista tiedonkeruumenetelmistä sekä kansainvälisiin verrokkeihin kohdistuvista havainnoista. Tässä artikkelissa kysymme:

- Miten datapaketit ja rajapinnat soveltuvat erilaisiin tutkimuksellisiin tarpeisiin? Mitä teknisiä mahdollisuuksia ja rajoitteita niillä on? Miten ne suhteutuvat käyttäjien tutkimuksellisiin tarpeisiin ja tekniseen osaamiseen?

Artikkelin teoreettisessa osuudessa avaamme käsitteet data, avointa dataa sisältävä datapaketti ja avoin ohjelmistorajapinta. Aineisto-osuudessa kuvaamme lyhyesti artikkelissa hyödynnetyn DAM-aineiston ja tutkijanäkökulman. Kansalliskirjaston digitaalisten aineiston tilastojen koostamisesta vastasi Tuula Pääkkönen. DAM-aineiston analyysissa hyödynnettiin sisällön analyysiä ja teemoittelua, joita toteutti Liisa Näpärä, ja datapakettien ja rajapintojen teknisten havaintojen tulkinnasta Erno Liukkonen. Analyysiosiossa luokittelemme tutkijoiden teknologiset ja digitaaliset tarpeet kolmeen liukuvarajaiseen kategoriaan. Sen jälkeen vertaamme ja reflektoimme Kansalliskirjaston datapakettien ja rajapintojen tutkijakäytön tilastoja. Tämän jälkeen siirrymme yksityiskohtaisemmin siihen, mitä tarpeita tutkijat esittivät näille haastatteluissa. Artikkelin loppupuoli keskittyy datapakettien ja rajapintojen vertailuun sekä niistä tehtyihin havaintoihin vertailuaineistosta. Tarkastelemme erityisesti datapakettien ja rajapintojen mahdollisuuksia ja rajoitteita tutkijakäytössä. Huomioimme niiden erilai-

set tekniset ominaisuudet ja organisaatioiden asettamat reunaehdot niiden käytölle sekä tutkijoiden teknologiset ja digitaaliset tarpeet tutkimuksen toteutuksessa. Kokoamme ne lopuksi yhteenvetotaulukkoon.

Datapaketit ja rajapinnat tarjoavat avointa dataa

Pyrkimys avoimuuteen on korostunut viime vuosina julkisten organisaatioiden toiminnassa ja tieteellisessä tutkimuksessa niin Suomessa kuin kansainvälisesti. Kulttuuriperintöorganisaatiot tavoittelevat kokoelmiensa muuttamista dataksi (*collections as data*). Avoin data, avoin rajapinta ja avoin lähdekoodi ovat termejä, jotka nousevat usein esiin puhuttaessa avoimuutta edistävistä digitaalisista menetelmistä. Uudet digitaaliset menetelmät mahdollistavat datan julkaisun ja sen hyödyntämisen monipuolisesti, mutta ne muodostavat myös haasteita organisaatioille. (Candela et al., 2020; Sugimoto, 2017, 315–316.)

Kansalliskirjasto pyrkii edistämään laajasti tieteen avoimuutta, ja avoimuus on yksi sen toiminnan kulmakivistä. Kansalliskirjastossa avoimuudella tarkoitetaan periaatetta, jossa sitoudutaan avoimen tieteen periaatteisiin, ideoita vastaanotetaan ja edistetään vapaasti sekä toiminnoissa että toimintakulttuurissa (Kansalliskirjasto, 2021). Avoimuus näkyy esimerkiksi Kansalliskirjaston digitaalisten ihmistieteiden politiikassa (Kansalliskirjasto, 2016) ja avoimessa Kansalliskirjasto-politiikassa (Kansalliskirjasto, 2017). Näiden perusteella on myös luotu toimintasuunnitelmia, joissa aktiivisesti pyritään avoimuuteen läpi organisaation ja palvelujen. Avoimuus ei siis rajoitu pelkästään aineistojen avoimuuteen, mutta se on yksi keskeisistä avoimuuteen liittyvistä asioista. Kun aineistoja ja kokoelmia muutetaan digitaalisiksi ja edelleen rakenteistetuksi dataksi, edistetään usein avoimuutta. Datan avoimeen käyttöön heijastuvat FAIR-periaatteet, joita ovat löydettävyys (Findable), saavutettavuus (Accessible), yhteentoimivuus (Interoperable) ja uudelleenkäytettävyys (Reusable) (EU-neuvoston linjaus, 2016; Padilla et al., 2017).

Datapaketit

Datasta puhutaan eri konteksteissa ja eri tavoin. Tässä artikkelissa datalla tarkoitetaan digitaalista raakamateriaalia, jota voidaan käyttää tutkimuksessa. Data on koneluettava ja rakenteistettu digitaalisen aineiston muoto, jota tarjotaan käyttäjille. (Borgman, 2020, 994; Schöch, 2013.) Kun puhutaan muusta kuin rakenteistusta digitaalisesta datasta, artikkelissa puhutaan aineistoista, joka voi olla joko digitaalista tai analogista. Datana tarjottujen

aineistojen ja kokoelmien käyttö on usein monipuolisempaa kuin analogisten eli ei-digitaalisten aineistojen käyttö. Dataan on sovellettavissa erilaisia laskennallisia menetelmiä, kuten tekstianalyysi- ja kuva-analyysimenetelmiä. Niiden avulla tutkijoiden on mahdollista saada tutkimukseensa tarpeellista tietoa isoistakin datamassoista. (Ames, 2021, 1; Candela et al., 2020; Padilla et al., 2017.)

Datapaketeilla tarkoitetaan Kansalliskirjaston digitoiduista sanomalehdistä tarjoamaa dataa. Ne ovat ladattavissa digitoitujen aineistojen (Digin) avoimen datan sivuilta[1]. Lisäksi muuta dataa on tarjolla Kansalliskirjaston datakatalogista[2]. Datapaketeissa on sovittu dataformaatti tarjottaville tiedostoille kuin myös tietty rakenne, johon datapaketissa tarjottavat tiedostot on järjestetty. Kansalliskirjaston datapaketit sisältävät lehtiaineistoja esimerkiksi TIFF-, JPG-, ALTO XML- ja METS-muodossa. Kansalliskirjaston datapakettien laadinnassa on hyödynnetty *Open Knowledge Internationalin* luomaa "Data Package" -määritystä, jossa määritellään esimerkiksi, millaista yleistä metatietoa datapaketin tulee sisältää. Metatiedon tulee määrityksen mukaan sisältää tieto datapaketin nimestä, dataan liittyvistä lisensseistä ja tarkempaa tietoa datasta, kuten sen sijainnista datapaketin sisällä. Laadittujen Data Package -määritysten tulkitsemiseen on tehty useille eri ohjelmointikielille apuohjelmistoja (Candela et al., 2020, 8).

Avoin ohjelmistorajapinta

Rajapinnat ovat tapa tarjota digitaalista dataa rakenteisella, erikseen sovitulla tavalla. Niiden lähtökohtana on mahdollistaa tietojen siirto järjestelmästä tai muodosta toiseen. Tieto voi olla esimerkiksi dataa tai metadataa (Koster & Woutersen-Windhouwer, 2018). Rajapinta toimii yleensä sovellusten välissä, joten voidaan käyttää myös tarkempaa käsitettä ohjelmointirajapinta (Application Programming Interface, API). Se määrittelee, kuinka sovellus tarjoaa tietoja tai palveluita muille sovelluksille tai tietojärjestelmille.[3] Ohjelmointirajapinta voi myös olla datarajapinta, joka mahdollistaa pelkästään datan lukemisen toiseen sovellukseen, tai toiminnallinen rajapinta, jonka välityksellä voi esimerkiksi muuttaa sovelluksen tietoja. Esimerkkejä ohjelmointirajapinnoista ovat OAI-PMH, IIIF ja REST:n ylitse tarjottava JSON. OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protokollaa käytetään kansalliskirjastoissa ja muissa muistiorganisaatioissa metadatan haravointiin tutkijoille ja muille organisaatioille (Freire et al., 2017, 221).

1 Digin avoimen datan sivu (`https://digi.kansalliskirjasto.fi/opendata`)
2 Datakatalogi `https://data.nationallibrary.fi/`
3 Termipankki (`https://termipankki.fi/tepa/fi/`)

IIIF (International Image Interoperability Framework) -rajapintoja käytetään muistiorganisaatioissa kuva-aineistojen ja niihin liittyvien metatietojen välittämiseen (Raemy & Schneider, 2019, 9). REST on puolestaan arkkitehtuurityyli, jonka määrittelemien rajoitteiden avulla ohjataan rajapintojen suunnittelua ja kehitystä[4] (Verborgh et al., 2015).

Ollakseen avoin rajapinnan täytyy täyttää tietyt ehdot. Avoimen rajapinnan kaikki ominaisuudet ovat julkisia, eikä sen käyttöön liity rajoittavia ehtoja. Sen tulee olla avoimesti dokumentoitu, otettavissa käyttöön ilman ylläpitäjän tai järjestelmätoimittajan toimia, ja testattavissa, kuten kehittäjäyhteisön yhteisesti luomalla Avoin Rajapinta -sivustolla[5] määritelty. Jatkossa käytetään lyhyesti sanaa rajapinta tarkoittamaan avointa ohjelmointirajapintaa. Tosin avoimenkin rajapinnan kohdalla tekijänoikeusrajoitukset voivat rajoittaa käyttöliittymissä tai rajapintojen kautta saatavissa olevaa dataa, koska kaikki digitaalinen sisältö ei ole automaattisesti saatavilla datana tutkimuskäyttöön tai muuhunkaan käyttöön (Borgman, 2020, 993). Lisäksi kaikkia rajapintoja ei ole tarkoitettu julkiseen käyttöön, koska niitä hyödynnetään esimerkiksi kirjastojen tarjoamissa verkkosivuilta toimivissa palveluissa, joiden avulla esitetään dataa käyttäjille. Osa rajapinnoista on taas tarkoitettu tutkijoille ja muille käyttäjille, jotka haluavat suorittaa erilaisia rajapintakyselyjä (Mansaré, 2018).

Rajapinnat toimivat alustarajaresurssina (*platform boundary resource*). Rajapintojen ominaisuuksien avulla alustoja kehitetään kolmansien osapuolten toimesta, eikä alkuperäistä kehittäjää tarvita. He eivät saa työstään useinkaan kompensaatiota, mutta heidän toimintansa on olennaista resurssiekosysteemin kehittymiselle. Ghazawnehin ja Henfridssonin artikkelissa rajaresurssimallissa resursoinnilla tarkoitetaan prosessia, jossa alusta monipuolistuu ja laajenee. Prosessissa alustalle kehitetään uusia rajaresursseja, jotka mahdollistavat ulkopuolisten kehittäjien laajentavan alustan ominaisuuksia luomalla sovelluksia, jotka hyödyntävät uusia rajaresursseja. Tarve rajaresurssien kehittämiselle voi ilmetä sisäisesti, kun havaitaan etteivät nykyiset resurssit mahdollista alustan suotuisaa kehittymistä. Myös ulkopuolisilta kehittäjiltä tulevat toiveet voivat aiheuttaa tarpeen kehittää uusia rajaresursseja (Ghazawneh & Henfridsson, 2013, 174–177). Rajapintojen tarjoaminen tutkijoille mahdollistaa heille laajemman ja joustavamman pääsyn dataan kuin eri verkkosivuilta toimivien palveluiden kautta on mahdollista. Rajapintojen kautta on mahdollista suorittaa esimerkiksi eri organisaatioiden tarjoamien datojen yhdistämistä ja vertailua, mikä ei tavalli-

4 Working with JSON `https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Objects/JSON`

5 Avoin rajapinta `http://avoinrajapinta.fi/`

sesti ole mahdollista verkkopalveluiden kautta. Käyttäjät määrittelevät tarkasti, millaisia hakuehtoja rajapintakyselyissä käytetään, ja he voivat rajata tarkasti, mitä dataa ja missä muodossa kyselyn tuloksina palautuu. Rajapintoja voidaan käyttää joko suoraan ohjelmointiesimerkeillä, muulla dokumentaatiolla tai yleisempien käsitteiden kautta keräämällä laajempi käsitys rajapinnan tarkoituksesta. (Candela et al., 2020, 6; Meng, 2018; Sugimoto, 2017, 325.) Niiden avulla saadaan esimerkiksi automatisoitua tehtäviä ja tehtyä dataan tarkistuksia, jotka käsin voisivat olla erittäin aikaa vieviä.

Artikkelin aineisto

Artikkelin aineisto koostuu Kansalliskirjaston omista datapakettien ja rajapintojen käyttöä kuvaavista tilastoista, käyttäjälähtöisen tiedon keruusta ja kansainvälisestä vertailuaineistosta. Aineiston avulla voimme tarkastella sekä tutkijoiden kokemuksia ja tarpeita että heille tarjottavia datan käytön ratkaisuja mahdollisimman monipuolisesti yhden käyttäjäryhmätapauksen avulla (Gagnon, 2010). Tutkijat ovat strategisesti tärkeä kohderyhmä tieteelliselle kirjastolle, joka tarjoaa tutkimuskirjallisuuden lisäksi myös tutkimuksen lähdemateriaalia (Kansalliskirjasto, 2021). Tutkijat ovat myös tilastollisesti aktiivisia Digi-palvelun[6] käyttäjiä. Vuonna 2019 tehdyssä palvelun käyttäjäkyselyssä käyttäjät ilmoittivat aineistojen käyttötarkoituksekseen sukututkimuksen (23 %), tutkimuksen yksityisiin tarpeisiin (18 %), selailun (14 %), tieteellisen tutkimuksen (13 %), tutkimuksen kirjaan/tietokirjaan (11 %), tutkimuksen oman organisaation/yrityksen tarpeisiin (10 %), korkeakouluopintoihin (7 %) ja muuta käyttöä (4 %). Tämän kyselyn jälkeen jäimme pohtimaan tutkijakäyttäjien luonnetta ja tavoitteita palvelujen kehittämiseksi eteenpäin.

Vuonna 2020 DAM-hankkeessa Kansalliskirjaston digitaalisia aineistoja käyttäneitä tutkijoita lähestyttiin kyselyllä ja haastatteluilla. Kyselyyn osallistui 130 vastaajaa, ja tyypillinen vastaaja oli Helsingin yliopistossa historian ja arkeologian alalla työskentelevä väitöskirjatutkija (Näpärä & Lilja, 2021). Lisäksi hankkeessa haastateltiin 18 tutkijaa 15 haastattelussa. Kaksi haastatteluista oli ryhmähaastatteluja. Lähes kaikki haastatellut tavoitettiin kyselyn avulla. Haastateltavien määrää voi pitää kohtuullisena ottaen huomioon kyselyyn vastanneiden 45 henkilön suostumuksen jatkoyhteydenottoihin. Tähän peilattuna siis hieman alle kolmannes heistä tavoitettiin haastattelua varten. Kyselyn luotettavuutta ja tutkijoiden tavoitettavuutta suhteessa Kansalliskir-

6 Digi https://digi.kansalliskirjasto.fi/

jaston digitaalisten aineistojen käyttäjiin arvioitiin Liisa Näpärän ja Johanna Liljan (2021) artikkelissa.

Lisäksi tiedonkeruussa hyödynnettiin kansainvälisten toimijoiden haastatteluja ja havainnointia. Usein tästä puhutaan vertaisoppimisena eli benchmarkkauksena, jota voidaan yleisesti hyödyntää esimerkiksi palveluiden, prosessien ja tuotteiden tunnistamisessa ja niitä vastaavien omien palveluiden, prosessien ja tuotteiden kehittämiseen (Bhutta & Huq, 1999). Oppia haettiin muiden kansalliskirjastojen library labeista, jotka ovat digitaalisten aineistojen ja niiden asiantuntijoiden sekä niiden käyttäjien yhteenliittymiä. Labien tavoitteena on digitaalisten aineistojen laaja ja avoin käyttö sekä laskennallisten menetelmien soveltaminen aineistojen analysointiin. Niissä kiinnitetään huomiota datan tarjontaan laadukkaasti eri käyttäjille sallien kokeilut, innovaatiot ja uuden synnyttämisen. (Mahey et al., 2019.)

Yhteensä seitsemän maan labien tai kehitteillä olevien labien edustajia haastateltiin, keskimäärin kaksi kustakin maasta. (Näpärä & Liukkonen, 2020.) Havainnot tehtiin puolestaan avoimesti saatavasta internetissä julkaistusta aineistosta, joka tuki haastatteluissa saatua tietoa. Havainnointiaineisto muodostui teknisten työkalujen dokumentaatiosta, datapakettien sisältöjen ja rajapintojen kuvauksista verkkosivuilla, joista osaan viitataan suoraan tässä artikkelissa [hakasuluissa olevilla merkinnöillä] (ks. aineistoluettelo). Havainnointiaineisto, johon tässä keskitytään, avasi erilaisia teknisiä ratkaisuja, joita oli toteutettu datan tarjonnassa. Näin saatiin digitaalisista ympäristöistä monipuolisesti ymmärrystä aiheesta, ja oman toiminnan kehittämiseen tarvittavaa teknistä lisätietoa yhteen datan käyttäjäryhmätapaukseen keskittyen (Haanpää et al., 2014, 291; Thomas & Myers, 2015).

Datan käyttö ja digitaalisten aineistojen käyttäjät

Tässä luvussa kuvataan, miten Kansalliskirjaston Digissä olevaa digitoitujen aineistojen dataa on käytetty hyödyntäen sekä rajapintoja että datapaketteja. Lisäksi luvussa jaotellaan digitaalisten aineistojen käyttäjät kolmeen kategoriaan DAM-aineiston perusteella.

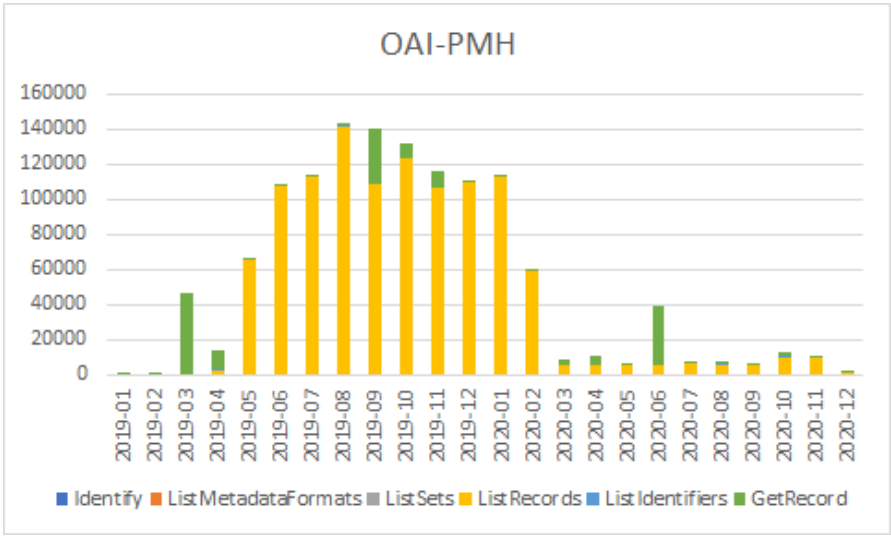
Tarjolla olevaan datan määrään ja sen käytön mahdollisuuksiin verrattuna DAM-aineistossa datapakettien ja rajapintojen käyttö näkyi odotettua vähemmän. Datapakettien ja rajapintojen käytöstä ei kysytty kyselyssä erikseen, mutta kyselyn 130 vastaajasta vain 16 (12,3 %) oli hyödyntänyt datakatalogia, joka on yksi reitti datapaketteihin ja rajapintoihin. Käyttäjät ovat voineet päätyä rajapintojen ja datapakettien pariin myös suorilla linkeillä. Lisäksi, jos he ovat käyttäneet niitä pidempään, ei yhteys datakatalogiin ole välttämättä kovin

selvä. Tämä heijastui myös haastatteluissa kuulluissa tutkijoiden kokemuksissa.

Vertailun vuoksi kyselyssä vastausvaihtoehtona ollutta Finna APIa (Finna-hakupalvelun rajapinta)[7] oli vastannut käyttäneensä 45 (37 %). On kuitenkin oletettavaa, että osa vastaajista on sekoittanut sen Finna-hakupalveluun, joka yhdistää eri kirjastojen, arkistojen, museoiden ja muiden toimijoiden aineistotietoja yhteen paikkaan. Haastatteluissa kävi ilmi, etteivät kaikki haasteltavat tienneet rajapinnan olemassaolosta tai merkityksestä. Hakupalveluna Finna oli kuitenkin kerätyn aineiston mukaan varsin tunnettu ja käytetty, vaikka sen käytöstä ei erikseen kysytty kyselyssä tai haastatteluissa. Finnan tunnettuus sekä käyttäjämäärät näkyvät myös tilastoissa, joiden mukaan vuonna 2020 kaikissa Finna.fi-näkymissä oli 44,2 miljoonaa käyntikertaa, kun taas tässä fokuksena toimivassa Digi-palvelussa oli 13,2 miljoonaa sisältösivun latausta (Kansalliskirjaston vuosi 2020). Hakupalveluiden käyttömäärät sisältävät selauskäytön sekä palveluiden sisäisiin rajapintoihin kohdistuvat pyynnöt, koska käyttäjien palveluissa tekemät haut ja palveluiden sivujen käyttö aiheuttaa pyyntöjä palveluiden sisäisiin rajapintoihin.

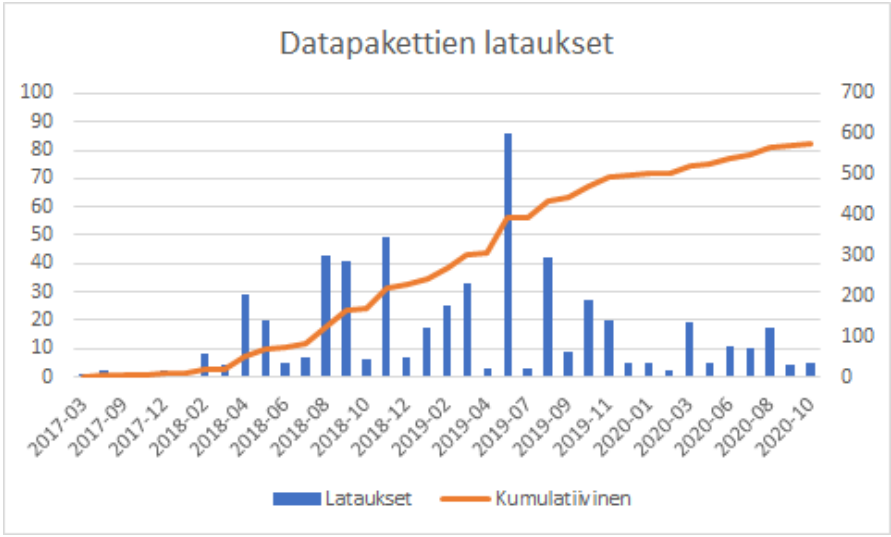
Kansalliskirjastojen rajapintojen hyödyntäminen on mahdollista sekä palveluiden sisäiseen että avoimeen käyttöön tarkoitettujen rajapintojen kautta. Avoimeksi tarkoitettuja rajapintoja hyödyntävät esimerkiksi tutkijat ja dataa haravoivat organisaatiot, kuten Europeana[8] (eurooppalaisen kulttuuriperinnön portaali), joka tarjoaa pääsyn eurooppalaisista kirjastoista, museoista, arkistoista ja muista muistiorganisaatiosta haravoituun avoimeen dataan (Edmond & Garnett, 2015, 288); [europeana mission]. Haravointia suorittavat organisaatiot muodostavat merkittävän osan avoimeen käyttöön tarkoitettujen rajapintojen käytöstä. Seuraavassa kuvassa näkyy OAI-PMH rajapinnan käyttö vuonna 2020. Käytössä näkyy myös Europeanaan vuonna 2020 haravoitu data. Muu käyttö on alle 10.000 kyselyä kuukaudessa, eikä esimerkiksi tutkijakäyttöä erotella muusta käytöstä.

7 https://api.finna.fi/
8 Europeana https://www.europeana.eu



Kuva 1: OAI-PMH rajapinnan käyttötilasto.

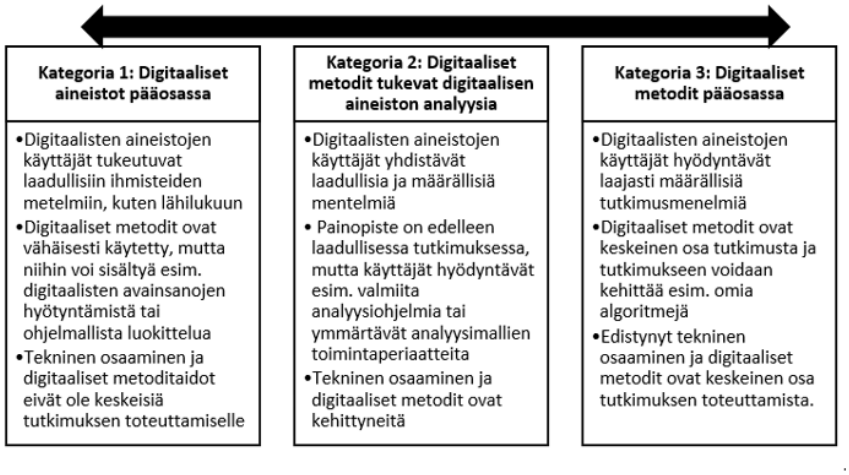
Kansalliskirjaston datapakettien lataus on määrällisesti rajapintakäyttöä vähäisempää. Datapaketteja oli ladattu DAM-hankkeessa kerätyn kyselyn tekohetkellä kumulatiivisesti yli 500 kertaa vuodesta 2017 alkaen. Datapakettien lataus on vakiintunut noin 10–20 paketin lataukseen kuukaudessa. Alla oleva kaavio näyttää loppuun asti suoritetut lataukset eri kuukausina, datapakettien lataustoiminnon avaamisesta vuoden 2020 loppupuolelle.



Kuva 2: Digitaalisten aineistojen datapakettien lataustilasto.

Datapakettien latausmäärät perustuvat IP-osoitteisiin, mutta lataajien tarkempi yksilöinti ei ole mahdollista, vaikka lataamisen yhteydessä on määritelty pakolliseksi käyttötarkoituskysely, johon voi kuitenkin vastata anonyymisti. Tilastoinnissa käytetyn IP-osoitteen takana voi myös olla useampia lataajia, mikä voi aiheuttaa tuloksiin epätarkkuutta. Lisäksi lataukset jakautuvat eri datapaketteihin epätasaisesti; eräinä vuosina on ladattu sanoma- ja aikakauslehtiin liittyviä datapaketteja, toisinaan taas eri tutkimusyhteistyöprojekteissa tuotettuja datapaketti, kuten vuonna 2021 uuden tekstintunnistuksen ground truth -paketti. Ground truth on opetusaineiston sisältävä datapaketti, joka muodostuu sanomalehtien ALTO-XML- ja kuvatiedostoista ja jota on käytetty tekstintunnistuksessa käytettävän mallin opetuksessa. Voikin olettaa, että käyttäjät eivät lataa isoja datapaketteja monta kertaa, vaan ne pysyvät käyttäjällä talletettuna useammankin vuoden, esimerkiksi koko tutkimusprosessin ajan tai useiden tutkimusten ajan. Datapaketteihin on tullut vain yksittäisiä palautteita, mikä viestii samasta.

DAM-aineiston sisällön analyysin perusteella tutkijat voidaan jakaa kolmenlaisiin digitaalisten aineistojen käyttäjiin heidän kokemustensa ja tutkimustarpeidensa perusteella. Taustalla on myös heidän teknologiset taitonsa ja digitaalinen osaamisensa, jonka kautta he usein perustelivat omaa digitaalisten aineistojen käyttötapaansa. Tutkijoiden kategorisointia on havainnollistettu seuraavassa kuvassa, jossa digitaaliset aineistot sisältävät myös muun kuin datamuotoiset sisällöt (Kuva 3: Tutkijoiden digitaalisten aineistojen käyttötavat). Kategorisointi ei arvota tutkimustaitoa.



Kuva 3: Tutkijoiden digitaalisten aineistojen käyttötavat.

Ensimmäisessä kategoriassa (jatkossa kategoria 1) on pääosin digitaalisiin aineistoihin ja laadullisiin menetelmiin tukeutuvat käyttäjät. Tutkimuksen toteuttamisen näkökulmasta tekninen osaaminen tai digitaaliset metoditaidot eivät ole olennaisia. Toisessa kategoriassa (kategoria 2) käyttäjät hyödyntävät digitaalisia metodeja tukemaan aineiston laadullista analyysia. Kolmannessa (kategoria 3) painopiste on digitaalisten metodien hyödyntämisessä ja tietokoneavusteisessa data-analyysissa, jolloin tekninen osaaminen ja digitaalinen metodiosaaminen on olennaista tutkimuksen toteuttamiseksi. DAM-aineiston perusteella tutkijoiden digitaalisen metodiosaamisen ja teknologian hyödyntämisen taitotarpeet vaihtelevat tutkimusasetelman mukaan. On mahdollista, että tutkimuksessa hyödynnetään edistyneempiä menetelmiä kuin yksittäisellä tutkijalla on. Tällöin hyödynnetään toisten tutkijoiden osaamista tutkimuksen toteuttamiseksi. Toisaalta tutkijan ei tarvitse hyödyntää koko osaamiskapasiteettiaan yhdessä tutkimuksessa. Sen vuoksi kuvassa on ylhäällä nuoli kuvaamassa sitä, ettei tutkijan tarvitse pysyä yhdessä kategoriassa, vaan tutkimuksellisten tarpeiden vaihtuessa hän voi liikkua kategoriasta toiseen. Hieman vastaavalla tavalla Sarah Amesin (2021) Skotlannin kansalliskirjaston datapalveluiden kontekstista kirjoittamassa artikkelissa käyttäjät kategorisoitiin kolmeen luokkaan: aloittelevat, edistyneet ja näiden kahden väliin jäävät tutkijat. Heidän kuitenkin luokiteltiin enemmän pelkän teknologisen osaamisensa perusteella, mutta myös Amesin artikkelissa koettiin tärkeäksi, että erilaisia datapalveluiden käyttäjiä palveltaisiin heidän tarpeensa ja osaamisensa huomioiden. Keskitason osaamisella varustetut tutkijat ovat Amesin mukaan haastavimmin palveltava käyttäjäryhmä, koska heillä on rajoitetut tekniset taidot, mutta he ymmärtävät erilaisten lähestymistapojen ja formaattien merkityksen tutkimusasetelmalle. (Ames, 2021, 4.)

DAM-aineiston analyysin perusteella enemmistö tutkijoista näyttää sijoittuvan teknisiltä taidoiltaan aloitteleviin ja ei-digitaalisia metodeja tutkimuksessaan hyödyntävään kategoriaan (Kuvassa 1: kategoriat 1 ja 2). Pääsääntöisesti DAM-aineiston maininnat datapaketeista ja rajapinnoista perustuvat muutamien teknisiltä taidoiltaan edistyneiden tutkijoiden näkemyksiin, koska heillä oli edistynyttä teknistä osaamista (erityisesti kategoria 3). Toisaalta kokonaisuudessaan tutkijoiden tekninen osaaminen digitaalisten aineistojen käyttäjinä ja digitaalisten metodien hyödyntäjinä asettui hyvin laajalle skaalalle aina digitaalisten materiaalien tulostajista datan tiedonlouhijoihin, eikä tarkkoja rajoja eri käyttäjäryhmien välille voi asettaa tutkimusasetelman vaikuttaessa tarvittaviin teknologisiin ja digitaalisiin taitoihin. Lisäksi datan avoimuus ja käyttöoikeudet ymmärrettiin eri käyttäjäryhmissä vaihtelevasti. Tämä näkyy esimerkiksi kyselyssä ja haastatteluissa toistuvina

vaatimuksina avata dataa nykyistä enemmän, vaikka datan käyttöoikeudet eivät tätä aina mahdollista.

Datapaketit tutkijakäytössä

Tutkijoille tehdyissä haastatteluissa Kansalliskirjaston tarjoamia datapaketteja pidettiin pääsääntöisesti hyvinä, mutta joitain kehitysehdotuksia annettiin. Muutamassa haastattelussa toivottiin esimerkiksi erikokoisia paketteja. Esimerkiksi pienikokoisten näytepakettien avulla dataan pääsee tutustumaan ennen varsinaisen suuremman datapaketin lataamista (Ames, 2021, 9). Myös opetuskäyttöön tarkoitettuja pienempiä paketteja toivottiin, kuten Luxemburgin kansalliskirjaston datasivuilla [Luxemburg data].

DAM-aineiston perusteella on tarpeen kiinnittää nykyistä enemmän huomiota datapaketteja tarjoavalta verkkosivulta saatavaan tietoon. Amesin mukaan sekä vertailuaineistoon perustuvien havaintojen perusteella datapakettien sisällöstä on tarpeen kertoa yksityiskohtaisesti verkkosivulla, josta paketit ovat ladattavissa. Datapaketteihin on tarpeen liittää yksilöivät tunnisteet, joihin tutkija voi tutkimuksessaan viitata. Jokaisessa datapaketissa pitää myös olla tieto siitä, milloin sitä on viimeksi päivitetty ja miten usein sen sisältöön kohdistuu päivityksiä. (Ames, 2021, 4–7.) Tämä helpottaa tutkijan pohdintaa siitä, mikä saatavilla olevista paketeista tai sisällöistä soveltuu parhaiten hänen käyttötarkoitukseensa. DAM-aineiston kokonaisuuden perusteella nykyistä pienempi datapaketin sisältämä historiallinen aikarajaus voi sopia tutkijan fokusoituun tiedontarpeeseen. Esimerkiksi laadullisesti suuntautuneet tutkijat, jotka kuitenkin osaavat teknisten taitojensa avulla hyödyntää datapaketteja (kategoria 2), voivat löytää tarvitsemansa kokonaisuuden pienemmästä määrästä dataa esimerkiksi melko vähäisten tai rutiiniluonteisten tietokoneavusteisten menetelmien hyödyntämisen jälkeen. Joskus taas on tarpeen lähteä haarukoimaan tutkimusaihetta isommasta massasta, jolloin suuremmat datapaketit palvelevat tutkimustarkoitusta paremmin. Erityisesti määrällisesti suuntautuneet ja teknisiltä taidoiltaan edistyneet tutkijat (kategoria 3) tarvitsevat isoja datamassoja, joista louhivat tutkimuskysymyksiin liittyviä asiayhteyksiä. Heille isompikokoiset datapaketit voivat olla hyvinkin soveltuvia.

Datapaketit ovat tarpeellisia monille tutkijoille, koska kaikilla ei ole välttämättä tarvittavaa osaamista tai halua ryhtyä käyttämään rajapintoja, jotka vaativat usein jonkin verran enemmän teknistä osaamista kuin valmiit datapaketit. Toisaalta kansainvälisessä tutkijoille tehdyssä kyselyssä havaittiin myös koodaustaitoja hallitsevien ja teknisiltä taidoiltaan edistyneiden tutkijoiden käyttävän valmiita datapaketteja. He ovat ladanneet datapaketit

suoraan verkkosivuilta tai joku heille tuttu henkilö, jolla on pääsy dataan, on toimittanut datapaketit. (Edmond & Garnett, 2015, 290–293.) Amesin (2021) mukaan datapakettien luonnissa täytyy huomioida tutkijoiden vaihtelevat tekniset valmiudet, ja millaisia työkaluja tutkijat ovat valmiita käyttämään data-analyysissa ja mitä vaatimuksia analyysitapa asettaa datapaketin sisällönn tiedostoformaateille. Vähemmän teknisesti orientoituneet tutkijat voivat esimerkiksi käyttää verkkosivuilta toimivia valmiita työkaluja (kategoria 2), jolloin heidän käyttöönsä voivat riittää pelkät tekstitiedostot, jotka sisältävät digitaalisen aineiston tekstin. Teknisiltä taidoiltaan edistyneemmät tutkijat (kategoria 3) hallitsevat puolestaan laajemmin erilaisten työkalujen käytön, ja heidän käyttöönsä voidaan tarjota dataa erilaisissa standardisoiduissa tiedostoformaateissa. (Ames, 2021, 4–7.)

Rajapinnat tutkijakäytössä

Rajapintojen kerrottiin DAM-aineistossa täyttävän niitä datan hankinnan aukkoja, joihin datapaketit eivät niinkään sovellu. Eräässä haastattelussa rajapintojen kautta tapahtuvaa aineistojen latausta pidettiin esimerkiksi parempana ja vähemmän työläänä vaihtoehtona kuin valmiita datapaketteja, koska siinä pystyy tarkasti määrittelemään esimerkiksi, mitkä sivukuvat (kuvia sivuista, joista esimerkiksi sanomalehdet muodostuvat) ja ALTO XML -tiedostot (raakatekstin sisältävä XML-skeema, joka kuvaa sivun ulkoasua) ladataan. Datan tarvitsijan ei tarvitse tällöin purkaa suuria datapaketteja ja selvittää, onko paketin sisällönkuvailu heidän tutkimusaiheensa kannalta riittävä. Rajapintojen avulla haastateltava pystyy rajaamaan datasta tarvitsemansa tiedostot tarkemmin kuin datapaketeista, jotka pitää ladata kokonaisina ja jotka voivat sisältää paljon tutkijalle tarpeetonta dataa.

Rajapintojen käytössä oli kuitenkin myös joitakin haasteita. Eräs haastateltu kertoi esimerkin tilanteesta, jossa hän haluaisi saada rajapinnan kautta palautettua sanomalehtiaineistoon liittyvän PDF-tiedoston linkin, jotta hän pääsisi käsiksi PDF-tiedostoon ja erottelemaan PDF-tiedostossa olevat kuvat luomallaan koodilla. Haasteena oli ollut se, että hän oli joutunut usein selvittämään PDF-tiedostojen linkit suoraan käyttämänsä Fenno-Ugrican verkkosivuilta[9]. Fenno-Ugrica on Kansalliskirjaston uralilaisilla kielillä painettujen julkaisujen digitaalinen kokoelma. Fennougrican ja Digin -palvelujen sisäisiä rajapintoja ei ole dokumentoitu julkiseen käyttöön tekijänoikeuksien vuoksi, joten niiden hyödyntäminen on haasteellista tutkijoille. Palveluiden rajapintoihin ei myöskään haluta ylimääräistä kuormitusta, joka voisi heikentää

9 Fenno-Ugrica aineisto `https://fennougrica.kansalliskirjasto.fi/`

palveluiden toimintaa. Dataan liittyvän PDF-tiedoston linkin selvittäminen on kuitenkin Digi-palvelun osalta mahdollista rajapintojen kautta ja datan sisältämien sivukuvien lataaminen jpg-formaatissa onnistuu myös rajapintojen kautta, jolloin kuvia ei tarvitse irrottaa PDF-tiedostosta.

Eräs haastateltava kaipasi selkeämpiä esimerkkejä, miten rajapintaa käytetään, koska hän on ensisijaisesti tutkija ihmistieteissä, ei tietojenkäsittelijä, vaikka DAM-aineistossa hän oli tulkittavissa teknisiltä taidoiltaan ja tarpeiltaan kategoriaan kolme. Go Sugimoton tutkimuksessa kansainvälisesti avoimien rajapintojen onkin havaittu olevan suunnattu ensisijaisesti tutkijoille, joilla on ohjelmointikokemusta, ja ohjelmistokehittäjille, jotka hallitsevat rajapintojen käytön. Sen sijaan ihmistieteilijöillä, joilla ei ole rajapintojen käyttöön vaadittavia ohjelmointitaitoja, rajapintojen käytössä on esiintynyt ongelmia. (Sugimoto, 2017, 325.) Ohjelmointitaitojen puute vähentääkin rajapintojen käyttöpotentiaalia tutkimuskäytössä, mikä on omiaan vaikeuttamaan monien humanistitaustaisten tutkijoiden työtä niiden parissa.

Datapakettien ja rajapintojen käytön mahdollisuudet ja rajoitteet

Datapakettien ja rajapintojen tutkijakäyttöön liittyy mahdollisuuksia ja rajoitteita. Osa näistä riippuu niitä tarjoavan organisaation tekemistä ratkaisuista esimerkiksi FAIR-periaatteiden osalta, ja osa on taas riippuvaisia käyttäjän teknisestä osaamisesta. Tässä luvussa on pohdittu, miten erilaiset datan tarjonnan tavat eroavat toisistaan. Luku koostuu neljästä alaluvusta. Kolme ensimmäistä alalukua kietoutuvat FAIR-periaatteisiin, keskittyen erityisesti saatavuuteen ja yhteentoimivuuteen (1. ja 3. alaluku). Kuvailu (2. alaluku) puolestaan liittyy enemmän FAIR-periaatteiden kaikkiin kohtiin. Luvun lopussa datapakettien ja rajapintojen mahdollisuudet ja rajoitteet tutkijakäytössä on koottu yksityiskohtaiseksi taulukoksi.

Käyttöliittymä ja lataussivu mahdollistavat datan saatavuuden

Data on löydettävissä ja saatavilla eri tavoin Kansalliskirjaston palveluista. Eri vaihtoehdot soveltuvat erilaisiin digitaalisiin ja teknologisiin tarpeisiin (kategoriat 2 ja 3). Rajapintakäytössä hyödynnetään erilaisia palvelun tarjoajien käyttöliittymiä ja tutkijoiden kehittämiä työkaluja, joiden kautta rajapintoja voidaan käyttää monipuolisemmin. Datapaketit puolestaan ladataan verkkosivun kautta ilman työkaluja tai käyttöliittymää.

Rajapintojen käyttöliittymien avulla kerätään haluttu data, jota voidaan edelleen käsitellä muissa työkaluissa (Edmond & Garnett, 2015, 294). Käyttöliittymiä on toteutettu esimerkiksi kuva-aineistojen rajapintojen hyödyntämiseen Hollannin ja Itävallan kansalliskirjastoissa [Hollanti rajapintatyökalu; Itävalta sacha]. Lisäksi on toteutettu erilaisia Jupyter-työkirjoja (Jupyter Notebook), joiden avulla on mahdollista ladata dataa rajapintojen kautta, suorittaa data-analyysia ja visualisoida analyysista saatuja tuloksia [glam workbench]. Käyttöliittymät helpottavat rajapintojen käyttöä, mutta eivät mahdollista yhtä monipuolista rajapintojen ominaisuuksien hyödyntämistä kuin tutkijoiden itse kehittämät rajapintatyökalut, jotka koodaustaitoiset tutkijat ovat luoneet myös muiden käytettäviksi (Puschmann & Ausserhofer, 2017, 153). Edmondin ja Garnettin (2015) tutkimuksessa osa haastatelluista tutkijoista koki valmiiden käyttöliittymien ja työkalujen rajoittavan rajapintojen ja niiden kautta ladatun aineiston tarjoamia käyttömahdollisuuksia. Ne tarjosivat apua datan käsittelyyn, mutta eivät tarjonneet ratkaisua kaikkiin haluttuihin käsittelytapoihin. (Edmond & Garnett, 2015, 294–295.)

Datapaketit ladataan käyttöön verkkosivustolta. Usein se on dataa tarjoavan oma verkkosivu, mutta myös Zenodoa ja Amazon AWS-pilvipalvelua voidaan käyttää (Ames, 2021, 5; Candela et al., 2020, 4). Syynä ulkoisen verkkopalvelun käyttöön on esimerkiksi se, että datapaketit vievät paljon tallennustilaa palvelimelta. Lisäksi niiden lataaminen vie datan tarjoajilta kaistaa, erityisesti tapauksissa, joissa useampia datapaketteja ladataan samanaikaisesti. Ulkopuolisten palveluiden käyttö datapakettien jakelussa vähentääkin kuormitusta kirjastojen palvelimille, ja mahdollistaa esimerkiksi Zenodon tapauksessa datapakettien laajemman löydettävyyden palvelun tarjoamien ominaisuuksien ansiosta. Ulkopuolisia palveluita käytettäessä täytyy kuitenkin huomioida, ovatko ne luotettavia ja pitkäaikaisia palveluntarjoajia, jotta datapaketit ovat ladattavissa myös pidemmällä aikavälillä luotettavasti tutkimuskäyttöön. (Candela et al., 2020, 4.)

Datan saatavuuteen voi samaan aikaan liittyä myös ehtoja tai teknisiä rajoitteita, jotka eivät mahdollista datan käyttöä kaikissa tapauksissa. Ghazawnehin ja Henfridssonin esittämässä rajapintaresurssimallissa alustan turvaamisella tarkoitetaan prosessia, jossa alustan omistaja kasvattaa alustan ja sen tarjoamien palveluiden hallintaa. Prosessin tarkoituksena on yleensä estää ulkopuolisia kehittäjiä luomasta sovelluksia, jotka voisivat vaikuttaa alustan toimintaan haitallisesti (Ghazawneh & Henfridsson, 2013, 177). Digi-palvelun osalta alustan turvaamista tapahtuu rajoittamalla käyttäjien pääsyä dataan rajapintojen ja datapakettien kautta. Myöskään kaikkea palvelun tarjoamien rajapintojen dokumentaatiota ei ole julkaistu ulkopuolisille kehittäjille.

Alustan suotuisan kehittymisen kannalta resurssointi- ja turvaamisprosessien välinen tasapaino on tärkeää (Ghazawneh & Henfridsson, 2013, 185–186).

Alustan turvaamisen tarkoituksena on datan tarkoituksenmukainen ja sopivasuhteinen käyttö käyttöehtojen mukaisesti esimerkiksi tieteelliseen tarkoitukseen. Esimerkiksi Kansalliskirjaston datapakettien kohdalla käyttäjä joutuu kuvaamaan datan suunniteltua käyttötarkoitusta. Australian Kansalliskirjaston digitaalisten aineistojen Trove-palvelun[10] rajapintojen käyttö vaatii myös käyttäjää rekisteröitymään sekä kertomaan siitä, kuinka on ajatellut käyttää rajapintoja [Trove apin käyttö]. Rekisteröinnillä voidaan ehkäistä organisaatioiden palvelimien kuormittamista, estää kilpailevaa toimintaa sekä tarkemmin yksilöidä, ketkä rajapintoja käyttävät, ja tarvittaessa asettaa rajoituksia rajapinnan käytölle. Rekisteröinti mahdollistaa myös organisaatiolle mahdollisuuden rakentaa kaupallista toimintaa rajapintojen ympärille ja periä maksuja rajapintojen käytöstä. Rekisteröinti ei kuitenkaan ole täysin ongelmatonta, sillä se voi karkottaa osan datan potentiaalisista käyttäjistä, riippuen siitä miten hankalana käyttäjät rekisteröitymisen kokevat. (Ghazawneh & Henfridsson, 2013; Puschmann & Ausserhofer, 2017, 150.) Monet avoimena pidetyistä rajapinnoista vaativat käytännössä käyttäjän rekisteröintiä, vaikka avoimen rajapinnan määritelmän mukaan käyttäjän ei tarvitsisi kysyä lupaa rajapinnan haltijalta. Tilanteissa rajapinnan avoimuutta voidaan kritisoida riippuen siitä, tapahtuuko rekisteröinnin hyväksyntä automaattisesti vai täytyykö käyttäjän kertoa enemmän käytön tarkoituksesta ennen rekisteröinnin hyväksymistä.

Kuvailu- ja dokumentointitiedot kertovat, mitä dataa on tarjolla ja miten se on saatavissa

Jotta data olisi FAIR-periaatteiden mukaisesti löydettävissä, saavutettavissa, yhteentoimivaa ja uudelleen käytettävää eri järjestelmissä ja sovelluksissa, datan on sisällettävä kuvailua ja dokumentaatiota. Datapakettien ja rajapintojen kuvailun tarkoituksena on, että käyttäjät tietävät, millaista dataa paketit sisältävät ja miten haluttua dataa saa ulos rajapintojen kautta jossakin formaatissa ja tietyltä aikaväliltä (Ames, 2021, 7). Erityisesti datapakettien kohdalla on olennaista kuvata, millaisesta datasta se muodostuu ja mitä mahdollisia puutteita dataan liittyy. Lisäksi on tarpeen kertoa, onko data alkuperäisessä raakamuodossa tai onko dataa käsitelty jollain tavalla. Datan hyödyntäjän näkökulmasta on myös tärkeää tietää, mihin tarkoitukseen datapaketti on laadittu, mihin tarkoituksiin datapakettia on jo hyödynnetty, kohdistuuko data-

10 Australian digitaalisten aineistojen verkkopalvelu `https://trove.nla.gov.au/`

pakettiin päivityksiä ja mihin datapakettia ei tulisi käyttää. (Gebru et al., 2018, 1–8.) Kuvailu ja dokumentointi auttavat tutkijaa siis löytämään tarvitsemansa datan analysoitavakseen.

Kuvailutieto kertoo dataan kohdistuneista muutoksista, jolloin puhutaan versioinnista. DAM-hankkeessa tehdyn tarkastelun perusteella dataa tarjoavat kansalliskirjastot tarjoavat ainoastaan viimeisimmän version datapaketeista, vaikka datasta luodaan yleensä useampia varmuuskopioita pidemmällä aikavälillä ja tarjolla voisi olla useita versioita. Myös rajapintojen kautta saatava data rajoittuu yleensä viimeisimpään versioon, eikä aikaisempia versioita ole mahdollista ladata (Vander Sande et al., 2014, 953). Datan uudet versiot muodostavat kuitenkin haasteen tutkijalle dataan tehtävän viittauksen ja tutkimuksen toistettavuuden kannalta.

Kansainvälisessä vertailuaineistossa havaitsimme, että datan dokumentaatiossa on jonkin verran vaihtelua ja puutteita eri organisaatioissa, erityisesti rajapintoihin ja niiden tarjoamiin käyttöesimerkkeihin liittyen. Osassa dataa tarjoavissa kirjastoista dokumentaatiota pitää etsiä esimerkiksi kirjaston GitHubin kautta, kuten Tanskan kansalliskirjaston OAI-PMH-rajapintojen dokumentaatiota [Tanska oai-pmh]. Joidenkin tapauksien kohdalla dokumentaatiota puolestaan tuodaan selkeästi esille rajapintoja käsittelevien verkkosivujen yhteydessä. Esimerkiksi Australian Trove on toteuttanut rajapintoja käsitteleville sivuilleen osuuden, jossa opastetaan rajapintojen käyttöön esimerkkien ja yksityiskohtien avulla [Trove api] (Edmond & Garnett, 2015, 289). Esimerkit kertovat esimerkiksi, kuinka rajapintoihin rakennetut kyselyt palauttavat dataa rajapinnan kautta. Lisäksi dokumentaation avulla voidaan esittää todellisia tutkijan käyttötapauksia tutkimuskysymysten muodostamisesta ja kuinka näihin kysymyksiin on saatu vastauksia rajapintoja hyödyntäen. Näin tutkijalle syntyy luottamusta siihen, että rajapintojen kautta saatavan datan avulla saadaan aikaan tuloksia. (Edmond & Garnett, 2015, 295.)

Joissakin tapauksissa huomasimme verrokkiaineistossa, että osa rajapinnoista oli kokeellisia ja tarkoitettu kirjastoissa eri projekteissa toteutettujen työkalujen sisäiseen käyttöön, jolloin tarkan dokumentaation toteuttaminen oli jäänyt tekemättä. Se voidaan kokea tarpeettomaksi tai dokumentaatiota ei tehdä pysyväksi. Esimerkiksi Tanskan library lab tuo sivullaan esiin projektien kokeellisuuden ja toteamuksen siitä, että projektien tuotokset, kuten työkalut ja rajapinnat, eivät ole pysyvästi saatavilla [Tanska lab]. Datan jatkokäytön vuoksi se on ongelmallista. Lisäksi heikon dokumentaation tason on havaittu vähentävän käyttäjien halukkuutta käyttää rajapintoja erityisesti niissä tapauksissa, joissa sama data on saatavilla muilla tavoin (Neumann et al., 2018, 11).

Valtaosa verrokkiaineistossa läpikäydyistä rajapintoihin liittyvistä dokumentaatioista oli kirjoitettu englanniksi. Ainoastaan Ranskan kansalliskirjaston [Ranska iiif-api] ja Berliinin valtionkirjaston [Berliini json-api] dokumentaatiot oli kirjoitettu paikallisilla kielillä. Paikallisella kielellä kirjoitettaessa voidaan pyrkiä palvelemaan tutkijoita asioissa, jotka voivat olla helpommin ymmärrettäviä tutkijoiden omalla äidinkielellä. Näin toimimalla voidaan kuitenkin hankaloittaa erikielisten käyttäjien mahdollisuuksia hyödyntää rajapintoja. Toisaalta samaan aikaan yleisesti rajapintojen dokumentaatiota täydentävät ulkopuoliset lähteet on kuitenkin yleensä kirjoitettu englanniksi, joten yhtenäisyyden vuoksi myös rajapintojen dokumentaatio olisi loogista olla saatavilla englanniksi.

Rajapintojen dokumentaation laadinnassa voidaan käyttää apuna työkaluja, jotka automatisoivat dokumentaation laadintaa, tai dokumentaatio voidaan luoda täysin manuaalisesti (Neumann et al., 2018, 11). Työkalujen avulla luodut dokumentaatiot sisältävät yleensä valmiita esimerkkikyselyitä, joita voi ajaa rajapintojen kautta, ja käyttäjä voi tehdä myös muutoksia esimerkkikyselyihin ja nähdä palautuuko rajapinnasta halutun kaltaista dataa. Dokumentaation sisältämät esimerkit mahdollistavat datan palauttamisen rajapintojen kautta useissa eri formaateissa, jolloin käyttäjän ei tarvitse lähteä selvittämään, miten muunnos johonkin tiettyyn formaattiin onnistuu, ja hänelle selviää suoraan dokumentaatiosta, mitä tiedostoformaatteja rajapinta tukee. Esimerkiksi Swagger-työkalu mahdollistaa laadukkaiden ja standardien mukaisten rajapintojen suunnittelun ja automaattisen dokumentaation laadinnan rajapintamäärittelyiden pohjalta [swagger api].

Kansainvälisesti rajapintojen käytön opastukseen on myös laadittu erilaisia ohjeita, jotka täydentävät dokumentaatiota ja sisältävät linkkejä myös muihin lähteisiin, joista on tutkijalle apua. Rajapintojen käytön opastamiseen on voitu laatia esimerkiksi blogikirjoituksia tai videoita tai järjestää erilaisia työpajoja. (Edmond & Garnett, 2015, 289–295.) Kirjastoissa voi myös olla asiantuntijoina henkilöitä, joiden omat projektit tuovat esiin kirjaston tarjoamien rajapintojen ominaisuudet ja mahdollisuudet. Esimerkiksi Australian Kansalliskirjaston Tim Sherratt on luonut useita projekteja (Sherratt, 2013), jotka hyödyntävät Trove-palvelun rajapintoja (Edmond & Garnett, 2015, 290, 295). Tämän lisäksi Sherratt on toiminut aktiivisesti lab-verkostossa ja tuonut esiin rajapintojen tarjoamia mahdollisuuksia ja niiden käyttöä esimerkiksi Jupyter-työkirjojen avulla (Sherratt, 2020).

Datan yhteentoimivuus eri järjestelmissä ja organisaatioiden välillä on keskeistä

Datan yhteentoimivuus eri järjestelmissä ja organisaatioiden välillä on keskeistä, jotta datan käyttö onnistuu sujuvasti erityisesti tilanteissa, joissa dataa on louhittu eri tavoin. Onnistuessaan yhteentoimivuus on mahdollisuus, mutta edelleen on tilanteita, joissa yhteentoimivuudessa on puutteita.

Rajapinnat mahdollistavat data-aineistojen yhdistämisen analyysia suorittaviin työkaluihin, koska työkalujen koodissa voidaan tarkasti määritellä, millainen kutsu rajapintaan suoritetaan, ja data-analyysi voidaan aloittaa välittömästi latauksen valmistuttua. Datapaketeissa voi ensin joutua lataamaan useampia erilaisia paketteja, perehtymään niiden sisältöihin sekä muuntamaan datan analyysityökalun käyttämiin tiedostoformaatteihin. Se voi vaatia tutkijalta enemmän aikaa ja perehtymistä siihen, miten hän saa rajattua datasta tietyn osan ja kuinka data muunnetaan tarvittavaan tiedostoformaattiin. Lisäksi datapakettien uusissa versioissa tutkijan täytyy perehtyä paketeissa tapahtuneisiin muutoksiin, jotka pakettien ylläpitäjän tulee dokumentoida. Dokumentaation mukana pitäminen eri versioissa varmistaa, että data on koko ajan käsiteltävissä analyysityökaluilla. Mahdolliset rajapintojen toteutukseen tulleet muutokset voivat aiheuttaa muutoksia myös analyysityökalun rajapintakutsuihin, joita täytyy tarvittaessa muuttaa, jotta työkalu säilyy toimivana.

Rajapintojen avulla datan muutoksien hallinta on datapaketteja helpompaa, koska muutokset datassa näkyvät jopa lähes reaaliaikaisesti. Rajapintaan kytketty analyysityökalu voi vertailla esimerkiksi eri aikaan suoritettuja rajapintakutsuja ja kohdistaa vertailun vain kutsujen välillä tapahtuneisiin muutoksiin. Rajapintojen avulla on myös mahdollista suorittaa eri organisaatioiden tarjoamien datan vertailua ja niiden yhdistämistä. Myös datapaketeissa tämä on mahdollista, mutta niiden sisällössä tai käytetyissä tiedostoformaateissa voi olla paljon eroja eri organisaatioiden välillä, mikä hankaloittaa niiden vertailua.

Datapakettien yhteentoimivuutta voidaan edistää käyttämällä yleisesti tunnettuja tiedostoformaatteja ja hyödyntämällä esimerkiksi Open Knowledge Internationalin luomaa "Data Package" -määritystä pakettien laadinnassa. Tietojen vertailun ja yhdistämisen avulla tutkija voi tehdä uusia havaintoja, jotka eivät yhtä datalähdettä käyttäen välttämättä olisi mahdollisia. Standardoitujen rajapintojen käyttö mahdollistaa sujuvan eri organisaatioiden datan yhdistämisen ja vertailun, ja ne voivat selkeyttää rajapintojen ylläpitoa organisaatiossa. Rajapintojen mahdolliset erot eri organisaatioiden välillä täytyy kuitenkin huomioida tutkijoiden omien ja projektien yhteydessä toteutettavien työkalujen toteutuksessa, mikä tarkoittaa tarkempaa perehtymistä

rajapintojen dokumentaatioon. Osa rajapinnoista on kehitetty esimerkiksi kokeellisissa projekteissa, eikä niiden pysyvyydestä ole takeita. Olisi kuitenkin suositeltavaa, että rajapintoja ylläpidetään pitkäjänteisesti ja ylläpitoa suorittavilla henkilöillä on riittävät tekniset taidot, jotta laadittujen esimerkkien ja dokumentaation tulkitseminen sekä rajapintaan tarvittavien muutoksien toteuttaminen on mahdollista.

Yhteenveto

Seuraavaan taulukkoon on koottu yksityiskohtaisemmin ja kokoavasti rajapintojen ja datapakettien mahdollisuuksia (taulukossa valkoisella pohjalla plusmerkein) ja rajoitteita (taulukossa harmaalla pohjalla miinusmerkein) käyttäjille dataa tuottavien organisaatioiden ja teknisten ominaisuuksien asettamien reunaehtojen puitteissa. Alustarajaresursseissa merkityksellistä on sekä käyttäjän että dataa tuottavan organisaation toiminta. Taulukon sisältö perustuu verrokkiaineistossa tehtyjen havaintojen analyysiin käytössä olevista rajapinnoista ja datapaketeista sekä artikkelien tekijöiden omaan kokemusperäiseen tietoon. Taulukon sisältö on luokiteltu havaintojen, analyysin ja kirjoittajien kokemuksien perusteella kuuteen ryhmään, joista jokaisesta on löydettävissä mahdollisuuksia ja rajoitteita sekä rajapintojen että datapakettien näkökulmista. Taulukon luokittelu tarkentui analyysin aikana, mutta moniin siinä olevista teemoista keskityttiin jo aineiston keruuvaiheessa.

Taulukko 1: Datapaketin ja rajapinnan käytön vertailua käyttäjän ja dataa tarjoavan organisaation näkökulmasta.

	Datapaketit	**Rajapinnat**
Kohderyhmä	Kategoriat 2 ja 3	Kategoria 3
Käyttöliittymä	+Ei erillisiä käyttöliittymiä, laadukkaasti toteutettu lataussivu selkeyttää käytön aloitusta +Voidaan jakaa myös ulkopuolisten palveluiden, kuten Zenodon kautta	+Mahdollistaa helpomman rajapintojen käytön esim. rajapintakysely muodostuu automaattisesti +Tarjoaa käyttöesimerkkejä käyttöliittymän kautta ennen ohjelmointivaihetta
	-Heikosti toteutettu lataussivu, jossa on vähän dokumentaatiota, vaikeuttaa käyttöä	- Rajapinnan ominaisuuksien hyödyntäminen ei yhtä monipuolista kuin itsekehitettyjen työkalujen kautta
Datan rajaaminen	+Sopii tilanteisiin, joissa paketoitu data on ennalta rajattu, esim. koko tutkimuksen kohteena olevaan ajanjaksoon	+Käyttäjä rajaa itse datansa Esim. tiedostoformaatti ja aikaväli voidaan usein määritellä

	-Sisältää usein tutkimuksen kannalta ylimääräistä dataa	-Voi vaatia perehtymistä dokumentaatioon, jotta tarkan rajapintakyselyn laadinta onnistuu
Dokumentaatio	+Helposti käyttäjän hahmotettavissa, ei vaadi pitkää perehtymistä, koska organisaation tuottamaa dokumentaatiota on vähän ja sen metadata on selkeää	+Dokumentaation laadinnassa voidaan hyödyntää työkaluja, kuten Swagger +Käyttäjä voi muokata ja suorittaa työkalujen avulla laadittuja esimerkkikyselyitä
	-Datapakettien tuottajat dokumentoivat muutokset, esim. kuvailutietojen päivittäminen vie aikaa - Dokumentaation, kuvailun laadintaan ei tavallisesti ole käytettävissä automatisoivia työkaluja	-Rajapintojen ylläpitäjät dokumentoivat muutokset: yksittäisten muutosten dokumentointi käy kuitenkin työkalujen avulla sujuvasti -Vaatii käyttäjältä tutustumista dokumentaatioon & rajapinnan toimintoihin. -Myös organisaatiossa täytyy olla ajantasainen tieto dokumentaation tasosta.
Yhteentoimivuus Tutkijan työkalut	+Standardien käyttö	+Standardien käyttö +Voidaan kytkeä osaksi datan analyysityökaluja
	-Voi joutua lataamaan erilaisia paketteja ja perehtymään sisältöön ja muuntamaan dataa työkalun käyttämään formaattiin.	-Rajapintoihin tulleet muutokset täytyy huomioida työkalun rajapintakutsuissa. Esimerkiksi kesken tutkimuksen voi joutua miettimään kutsujen muotoilua uudelleen.
Yhteentoimivuus organisaatioiden välillä	+Datapakettien yhdistäminen ja vertailu on mahdollista, jos rakenne ja sisältö eivät eroa liikaa toisistaan +Organisaatioissa voidaan rajata, tilastoida ja seurata käyttäjiä vaatimalla rekisteröitymistä	+ Standardoitujen rajapintojen käyttö mahdollistaa sujuvan eri organisaatioiden tarjoaman datan yhdistämisen ja vertailun +Rajapinnan käyttäjien rekisteröinti mahdollistaa tarkemman käytön seurannan ja tarvittaessa rajoittamisen.
	-Datapakettien sisältö ja rakenne voivat vaihdella, eri organisaatioiden tarjoaman datan yhdistäminen ja vertailu on vaikeampaa -Käyttäjän rekisteröitymisen vaatimus voi hankaloittaa käytön aloitusta	-Toteutuksessa voi olla eroja organisaatioiden välillä -Pysyvyydestä ei ole takeita erityisesti kokeellisten projektien tuloksena syntyneiden rajapintojen kohdalla -Käyttäjän rekisteröinti voi hankaloittaa käytön aloitusta

Ylläpito	+Ulkopuolisia palveluita käytettäessä osa ylläpidosta voidaan ulkoistaa dataa tuottavan organisaation ulkopuolelle	+Standardoidun rajapinnan käyttö voi selkeyttää ylläpitoa organisaatiossa
	-Organisaation täytyy ylläpitää datapaketteja ja niiden dokumentaatiota - Palvelimia täytyy ylläpitää, jos datapaketit jaetaan organisaation verkkosivujen kautta	-Rajapintoihin tulevat muutokset täytyy toteuttaa ja dokumentoida -Organisaation täytyy ylläpitää rajapintoja hyödyntäviä työkaluja ja palvelimia ja käyttäjän työkaluja, jotka hyödyntävät rajapintoja.

Johtopäätökset

Tutkimuskysymyksiimme vastaamme, että nykyiset datapaketit ja rajapinnat vastaavat erilaisten tutkijoiden tarpeisiin. Tutkijoiden datan käytön taidoissa on kuitenkin kehitystarpeita, ja kansainvälisiä kehityskulkuja seuraamalla ja tavoittelemalla FAIR-periaatteiden toteutumista voidaan vastata nykyistä paremmin tutkijoiden tarpeisiin datan käyttäjinä. Tutkijoiden erilaiset tarpeet ja kokemus datan käytöstä aiheuttavat kuitenkin haasteita datan tarjoajille ympäri maailmaa. Tutkijoita ohjaavat heidän taitonsa sekä erilaiset datan saatavuuteen vaikuttavat tekniset rajoitteet ja yleinen käytettävyys. Ne vaikuttavat halutun tutkimusaineistokokonaisuuden muodostamiseen hidastaen tai monimutkaistaen sitä. Tutkijoiden datan tarpeet ja tekninen osaaminen täytyykin huomioida siinä, millaista dataa ja minkä muotoista dataa tarjotaan valmiiksi kuratoituina datapaketteina tai rajapintojen kautta.

Rajapintojen ja datapakettien sujuva käyttö vaatii tutkijalta kokemusta ohjelmoinnista, eikä se vielä onnistu kaikilta tutkijoilta, vaikka heillä olisi kiinnostusta digitaalisten aineistojen ja datan monipuoliseen hyödyntämiseen tutkimusaineistona. Datan tarjoajien puolestaan täytyy huomioida, että dokumentaatio on riittävällä tasolla edistyneemmille datan tarvitsijoille, ja millä ohjelmointikielillä mahdolliset rajapintojen käyttöä opastavat esimerkit on toteutettu. Rajapintojen dokumentaatiota ja niissä olevia esimerkkejä pitäisi tulevaisuudessa parantaa, jotta tutkijoiden esittämiä ongelmia niiden tutkimuskäytöstä voidaan ratkaista. Lisäksi palveluiden sisäiseen käyttöön tarkoitettujen rajapintojen kohdalla täytyy kuitenkin harkita, halutaanko niiden dokumentaatio tuoda julkisesti esille ja miten varautua esimerkiksi suorituskykyvaateisiin, kun rajapinta julkaistaan käyttöön.

Tutkijoiden dataan liittyvien jatkuvasti kasvavien intressien vuoksi datan tarjontaan pitää jatkossakin luoda mahdollisuuksia eri tavoin erityisesti FAIR-periaatteet ja yleinen datan käsittelyn helppous huomioiden.

Mahdollisuuksien tarjoaminen ei kuitenkaan poista tekijänoikeuksiin liittyviä rajoitteita, jotka ovat oma monimutkainen aihealueensa. Datapaketit ja rajapinnat sopivat eri tilanteisiin ja niiden käyttöön tarvitaan erilaisia teknisiä valmiuksia. Eri tahojen kanssa tehtävällä yhteistyöllä voidaan alentaa rajapintojen, datapakettien ja datan käytön kynnyksiä noudattamalla yhdenmukaisia käytäntöjä sekä kommunikoimalla nykytilasta ja rajapintojen ja datapakettien jatkokehityksestä mahdollisimman pitkällä jänteellä.

Kiitokset

Digitaalinen avoin muisti -hanketta on rahoittanut Euroopan aluekehitysrahasto Vipuvoimaa EU:lta 2014–2020. Digitaalinen avoin muisti on Kansalliskirjaston datalähtöisten palvelujen kehityshanke, jossa suunniteltavat ja myöhemmin rakentuvat tutkimuksen palvelut ovat digitaalisten palveluratkaisujen ytimessä ja hyödyttävät kaikkia digitaalisen aineiston käyttäjiä.

Lähteet

Verrokkiaineisto

[Berliini json-api] https://sigel.staatsbibliothek-berlin.de/en/suche/json-api/
[europeana mission] https://pro.europeana.eu/about-us/mission
[glam workbench] https://glam-workbench.github.io/
[Itävalta sacha] https://labs.onb.ac.at/en/tool/sacha/
[Luxemburg data] https://data.bnl.lu/
[Ranska iiif-api] http://api.bnf.fr/api-iiif-de-recuperation-des-images-de-gallica
[swagger api] https://swagger.io/solutions/api-documentation/
[Trove api] https://trove.nla.gov.au/about/create-something
[Tanska oai-pmh] https://github.com/kb-dk/access-digital-objects/blob/master/oai-pmh.md
[Trove apin käyttö] https://trove.nla.gov.au/about/create-something/using-api

Verrokkiaineisto kerättiin useiden eri organisaatioiden tuottamasta datasta, ja tässä on listattu vain ne, joita käytettiin artikkelissa suorina lähteinä.

Kirjallisuus

Ames, S. (2021). Transparency, provenance and collections as data: the National Library of Scotland's Data Foundry. *Liber Quartely, 31*(1), 1–13. http://doi.org/10.18352/lq.10371

Bhutta, K. S., & Huq, F. (1999). Benchmarking – best practices: an integrated approach. *Benchmarking: An International Journal, 6*(3), 254–268. https://doi.org/10.1108/14635779910289261

Borgman, C. (2020). Whose text, whose mining, and to whose benefit? *Quantitative Science Studies, 1*(3), 993–1000. https://doi.org/10.1162/qss_a_00053

Candela, G., Dolores Sáez, M., Escobar Esteban, M., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science, 46*(5), 1–17. https://doi.org/10.1177/0165551520950246

Edmond, J., & Garnett, V. (2015). APIs and Researchers: The Emperor's New Clothes? *International Journal of Digital Curation, 10*(1), 287–297. https://doi.org/10.2218/ijdc.v10i1.369

EU-neuvoston linjaus (2016). https://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf. Viitattu 19.3.2021.

FAIR-periaatteet https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/. Viitattu 29.4.2021.

Freire, N., Robson, G., Howard, J. B., Manguinhas, H., & Isaac, A. (2017). Metadata Aggregation: Assessing the Application of IIIF and Sitemaps Within Cultural Heritage. Teoksessa J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (toim.), *Research and Advanced Technology for Digital Libraries* (s. 220–232). Springer. http://doi.org/10.1007/978-3-319-67008-9_18

Gagnon, Y. (2010). *The Case Study as Research Method*. Presses de l'Universite du Quebec.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2018). Datasheets for Datasets. ArXiv. https://arxiv.org/abs/1803.09010

Ghazawneh, A., & Henfridsson, O. (2013). Balancing platform control and external contribution in third-party development: the boundary resources model. *Information Systems Journal, 23*(2), 173–192. https://doi.org/10.1111/j.1365-2575.2012.00406.x

Haanpää, M., Hakkarainen, M., & Garcia-Rosell, J.-C. (2014). Etnografia kehittämisen välineenä. Teoksessa P. Hämeenaho, & E. Koskinen-Koivisto (toim.), *Moniulotteinen etnografia* (s. 287–307). Ethnos. http://urn.fi/URN:ISBN:978-952-68509-3-1

Kansalliskirjaston Digital Humanities -politiikka (2016). https://www.doria.fi/bitstream/handle/10024/130994/KK_DH_politiikka_18042016.pdf?sequence=2. Viitattu 26.4.2021.

Kansalliskirjasto. (2017). Avoin Kansalliskirjasto: politiikka ja toimenpideohjelma - Avoin Kansalliskirjasto - Yleinen sivusto. https://www.kiwi.fi/display/avoinkk/Avoin+Kansalliskirjasto%3A+politiikka+ja+toimenpideohjelma. Viitattu 29.4.2021.

Kansalliskirjasto (2021). Tehtävät ja strategia. https://www.kansalliskirjasto.fi/fi/tehtavat-ja-strategia. Viitattu 28.4.2021.

Koster, L., & Woutersen-Windhouwer, S. (2018). FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections. *The Code4Lib Journal*, (40).

Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., . . . Wilms, L., with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P., & Papaioannou, G. (2019). *Open a GLAMLab*. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23–27 September 2019.

Mansaré, M. (2019). *Finna-palvelun avoimen rajapinnan hyödyntäjät ja heille viestiminen* [opinnäytetyö]. https://urn.fi/URN:NBN:fi:amk-2019060314397

Meng, M., Steinhardt, S., & Schubert, A. (2018). Application Programming Interface Documentation: What Do Software Developers Want? *Journal of Technical Writing and Communication, 48*(3), 295–330. http://doi.org/10.1177/0047281617721853

Neumann, A., Laranjeiro, N., and Bernardino, J., (2018). An Analysis of Public REST Web Service APIs. *IEEE Transactions on Services Computing, 14*(4), 957–970. https://doi.org/10.1109/TSC.2018.2847344

Näpärä, L. (2020) Digitaalisten aineistojen tutkimusta tehdään enimmäkseen perinteisin menetelmin. Scripta Selecta 17.8.2020. https://blogs.helsinki.fi/scriptaselecta/2020/08/17/dam-kysely-digitaaliset-aineistot/. Viitattu 20.4.2020.

Näpärä, L. & Liukkonen, E. (2020). Report on the benchmarking interviews in the Digital Open Memory project. Zenodo. http://doi.org/10.5281/zenodo.4285836

Näpärä, L., & Lilja, J. (2021). Kansalliskirjaston aineistojen jatkokäyttö tutkijakyselyssä. *Informaatiotutkimus, 40*(1), 27–62. https://doi.org/10.23978/inf.99425

Padilla, T., Allen, L., Varner, S., Potvin, S., Russey Roke, E., & Frost, H. (2017). Call for Nominations: National Forum Attendance. https://collectionsasdata.github.io/nominations/. Viitattu 22.12.2020.

Puschmann, C., & Ausserhofer, J. (2017). Social data APIs: Origins, types, issues. Teoksessa M. T. Schäfer, & K. van Es (toim.), *The datafied society: Studying culture through data* (s. 147–154). Amsterdam University Press.

Raemy, J., & Schneider, R. (2019). Suggested measures for deploying IIIF in Swiss cultural heritage institutions. Zenodo. http://dx.doi.org/10.5281/zenodo.2640416

Schlicht, H. (2021). Open Access, Open Data, Open Software? Proprietary Tools and Their Restrictions. Teoksessa S. Schwandt (toim.), *Digital Methods in the Humanities - Challenges, Ideas, Perspectives* (s. 25–57). Bielefeld University Press.

Schwandt, S. (2021). Introduction – Digital humanities in practice. Teoksessa S. Schwandt (toim.), *Digital Methods in the Humanities - Challenges, Ideas, Perspectives* (s. 7–22). Bielefeld University Press.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities, 2*(3). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Sherratt, T. (2013). From portals to platforms – building new frameworks for user engagement. Presented at the LIANZA, Hamilton, New Zealand. Zenodo. http://doi.org/10.5281/zenodo.3563238

Sherratt, T., Wilms, L., & Lingstadt, K. (2020). LIBER Webinar: Setting Up A GLAM Workbench In Your Library. Zenodo. http://doi.org/10.5281/zenodo.3743193

Sugimoto G. (2017). Battle Without FAIR and Easy Data in Digital Humanities. Teoksessa E. Garoufallou, S. Virkus, R. Siatri, D. Koutsomiha (toim.), *Metadata and Semantic Research. MTSR 2017* (s. 315–326). Communications in Computer and Information Science, vol 755. Springer. https://doi.org/10.1007/978-3-319-70863-8_30

Thomas, G., & Myers, K. (2015). *The anatomy of the case study*. SAGE.

Vander Sande, M., Colpaert, P., De Nies, T., Mannens, E., & Van de Walle, R. (2014). Publish data as time consistent web API with provenance. Teoksessa *Proceedings of the 23rd International Conference on World Wide Web* (s. 953–958). Association for Computing Machinery. http://doi.org/10.1145/2567948.2579217

Verborgh R., Hooland, S. V., Cope, A.S., Chan, S., Mannens, E., & Walle, R. (2015). The Fallacy of the Multi-API Culture:Conceptual and Practical Benefits of Representational State Transfer (REST). *Journal of Documentation, 71*, 233–252. https://doi.org/10.1108/JD-07-2013-0098